DOCUMENT RESUME

ED 256 778 TM 850 116

AUTHOR	Humphreys, Lloyd G.
TITLE	Race Differences in Tested Intelligence: Important Socially, Obscure Causally. A Review ... of "Bias in Mental Testing", by Arthur R. Jensen.
INSTITUTION	National Academy of Education, Washington, D.C.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	81
NOTE	43p.
PUB TYPE	Information Analyses (070) -- Viewpoints (120) -- Collected Works - Serials (022)
JOURNAL CIT	Proceedings of the National Academy of Education; v7 1980
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Book Reviews; Ethnic Bias; Intelligence; Intelligence Differences; *Intelligence Tests; Racial Attitudes; *Racial Differences; *Test Bias
IDENTIFIERS	Jensen (Arthur R)

ABSTRACT

This document is a book review of "Bias in Mental Testing" by Arthur R. Jensen. Jensen discusses intelligence as a phenotypic construct. The problem of ethnic differences in phenotypic intelligence is emotionally charged, which makes rational consideration of the issues difficult. The reviewer disagrees with the author's predisposition to propose and stress genetic explanations for test bias, to the exclusion of other explanations, and disagrees with his account of the various tests of the Spearman hypothesis. The reviewer also disagrees with Jensen's test theory discussion and cites several statistical errors in the book. Jensen's use of first principal component and first principal factor without discrimination as an operational definition of the general factor is questioned. Despite the criticisms, the reviewer believes that the data presented are generally sound and that those that bear on test bias cannot easily be dismissed. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B, Bennett

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

BOOK REVIEW*

Lloyd G. Humphreys
University of Illinois

Bias in Mental Testing, by Arthur R. Jensen, New York: The Free Press.

xiii + 786 p. $29.95.

---

*Edited by John B. Carroll. Preparation of this review has been sponsored by the National Academy of Education under a grant from The Ford Foundation for support of activities of the Academy concerning public understanding of research on education. The opinions expressed are those of the author and do not necessarily represent the position of either the National Academy of Education or The Ford Foundation.

Lloyd G. Humphreys is Professor of Education and Psychology at the University of Illinois, Urbana-Champaign. A PhD of Stanford (1938), he was Assistant Director for Education at the National Science Foundation during 1970-71. His research has been focused on psychometrics, the theory of human abilities, and the effects of education.

Introduction

This is a large ambitious book. There are 740 pages of text, a glossary of four and one-half pages, 28 pages of references, and 32 pages of separate name and subject indexes. There are almost 800 references, many of which are discussed in considerable detail, and the information content per page of text is extremely high. Even the well qualified reader is overloaded with information. Since most of Professor Jensen's critics are not well qualified readers of psychometric data and theory, they will have to be very careful and serious readers indeed to understand the empirical bases for the author's conclusions. Jensen tries to help by teaching basic psychometrics in the early chapters.

This can also be an important book if the information is interpreted carefully. The reader must take seriously the author's intent to discuss bias in mental testing independently of the nature-nurture issue and interpret the data in that light. The reader must do this even though the author is not himself completely consistent in avoiding that issue. If the data presented are viewed, as they must be, as phenotypic correlates, the data describe a racial difference of significant though not of overwhelming social importance. This evaluation is independent of the causes of racial differences on ability tests, and the difference will not go

away by pretending it does not exist.

There is, however, a fundamental problem. Most readers of this book, whether inclined to criticize or to accept the author's conclusions, have a common definition of bias that arises from a shared folk definition of intelligence. The essential ingredient of the folk definition is that intelligence is a capacity of the organism. This capacity, furthermore, is innate. This second ingredient is as obvious in statements made by the most ardent environmentalists as in those by hereditarians, although the former group may avoid discussing individual differences in this supposedly fundamental biological capacity.

Many persons applying this common definition of intelligence will conclude that a case for freedom from bias has not been made. For them, the data presented by Jensen are essentially irrelevant. Others will conclude that mental tests are unbiased because the data will prove to their satisfaction that intelligence tests do measure capacity. In the context specified, i.e., starting from the folk definition, both conclusions are incorrect. Jensen's definition of bias is very different from the common definition. He states quite properly at the outset that he is discussing intelligence as a phenotypic construct. This is a stance that is incompatible with the folk

definition of intelligence and the common definition of bias, although he encourages confusion from time to time by slipping into the folk definition of intelligence himself.

For the environmentalist, an intelligence test is biased if it does not measure real intelligence. A measuring device that depends on learning cannot measure fundamental biological capacity in groups that have experienced segregation, social discrimination, and poverty. Therefore, intelligence tests are biased. For many hereditarians, on the other hand, the data presented support the acceptance of intelligence tests as measures of real intelligence and in consequence the data support a genetic interpretation of race and ethnic differences on those tests. In the light of acceptable scientific evidence, however, a belief in either biological equality or inequality must be considered an article of faith. In the light of the logic of science, moreover, the construct of intelligence as capacity, whether individual differences in the native ability are recognized or not, must be discarded. Intelligence defined as capacity can neither be measured nor estimated. Every construct in science must be defined in such a way that it can enter into a theory that leads to testable hypotheses.

Redefining capacity as genotype solves some problems associated with the use and interpreta-

tion of intelligence test scores since some of the popular connotations of capacity are lacking. Given good data in a given population the genotype can be estimated from the phenotype with a degree of random error dependent on the correlation between phenotype and genotype. In the language of capacity, however, estimation of phenotype is assumed. For example, it might be believed that a particular adult has had his or her height depressed by some earlier environmental circumstance. The hypothesis, in other words, is that the genotype is higher than the estimate obtained from the phenotypic measure. But how does one estimate height, the phenotype, from an unknown genotype? The best that can be done is to estimate the individual's phenotype from a composite of phenotypic measures obtained from relatives. There will necessarily be a good deal of variability about this estimate, since the correlation between the composite and individual heights will be substantially less than unity. Also, the usual logic of hypothesis testing cannot be applied to differences from these expectations.

The estimate of the standard deviation of the observed scores around the estimated score in a particular array is called a standard error, but it is not the standard deviation of a sampling distribution. It is misnamed. One cannot, for example, conclude that a score below the fifth or any other centile in an array is

probably not a deviation from the estimated mean of the array. The distribution of observations in bivariate distributions determines the size of the correlation. Without the observed extreme scores the correlation would be higher -- and there would still be scores below the fifth centile in each array. The situation is analogous to having available complete distribution of means of random samples. The extreme sample means are as much an integral part of the distribution and are as random as the ones close to the population mean.

The problem with the concept of capacity does not change materially if one moves from the individual to the group. The mean capacity of the group is unknowable. The same can be said for the mean genotype for human intelligence, at least in terms of presently available genetic knowledge and techniques. The only information available is that obtained from phenotypic measurements. Trends in these over time, if well documented, do constitute important information, but in human populations the attribution of a particular cause to such trends is very hazardous.

In the remainder of this review I shall first summarize the contents of the book. Next I shall concentrate on four chapters that contain the methodology advocated by the author and others for testing for bias and the data that

have resulted from the use of those methodologies. Finally, I shall discuss some issues somewhat peripheral to the problem of bias. They are related to the overall merit of the book; i.e., it is possible to agree with Professor Jensen that mental tests are not importantly biased, as biased is here defined, but fault him in other areas.

CONTENT SUMMARY

The Preface is an important part of the whole. It describes briefly Professor Jensen's intellectual history as it pertains to the problem of bias. The criticisms of intelligence tests in the forties based on their supposed bias against children from the lower socioeconomic classes and the attempt to construct a test that would reduce drastically the mean difference between the social classes are described. The course of the author's attitudes toward these efforts ranging from initial enthusiasm to disillusionment as he grew in methodological sophistication is also described.

The Preface is followed by two chapters which summarize criticisms of tests gleaned from the professional literature and from judicial decisions, respectively. The author remarks without comment that professional criticisms do not differ appreciably from those that appear in the popular press. I do not find this hard to

believe. A distressingly large number of persons trained in the social and behavioral sciences have taken positions on test bias more largely determined by sentiment than by data. The author does comment on the naiveté of some of the judicial statements in the opinions, but judges should be more pitied than censored. I base this on the quality of the testimony provided by the so-called experts. The judge who recently handed down an opinion in which he disregarded completely the expert testimony offered may have been wiser than his fellows (Grady, 1980).

The first definition of bias appears in Chapter 3. Bias is a property of a test that leads to errors of inference from test scores that are related to group membership. Error is also defined quantitatively. Thus over or underestimation of criterion scores in school, industry, or the military as a function of group membership is both an error and evidence for bias. If a predictor correlates more highly with group membership than with the criterion (a presumptively reliable criterion), there is also evidence for bias. A difference in mean scores on the predictor test does not in itself constitute evidence for bias because the inferences drawn from the test may not be in error. An inference about capacity from test scores, whether based on individuals or groups, is also an error of inference, but it is one that is not

quantifiable. It is appropriately deemed evidence for bias in the person drawing the inference. It is not bias in the test. Note that this is equally an error whether the inference affirms or rejects a particular estimate of capacity. The chapter also contains a rational analysis of the problems of blacks in their drive to attain equality and of the varying definitions of affirmative action that are currently in use.

The next five chapters contain elementary statistical and psychometric concepts and methodologies that are for the most part basic to the understanding of quantitative definitions of test bias. I am not able to judge how much of this is intelligible to an untrained reader, but it represents a valiant attempt. There are a number of isolated errors, but these do not detract from the exposition for the unsophisticated reader. Chapter 4 discusses the distribution of intelligence, Chapter 5 the varieties of mental test items, Chapter 6 the statistical and conceptual properties of correlations and factors, Chapter 7 reliability and stability of intelligence test scores, and Chapter 8 validity and correlates of those scores.

There are two important errors in this series of chapters. The decision to discuss and interpret the distribution of scores on intelligence tests is the first. The author correctly

points out that the shape of the distribution of scores on any psychological test is a joint function of the difficulty (or popularity in preference tests) levels of the items and their intercorrelations, but then proceeds to disregard the implications for interpretation. These properties within wide limits are under the control of the test constructer. The apparent normality--it is rarely tested--of distributions of test scores in many samples is a combination of lucky accident and common test construction practices. Even when something very close to normality as evaluated statistically is obtained, it cannot be used to support a polygenic hypothesis for the genetic basis for intelligence. For example, there is an anomaly in the near normal distribution of mental ages obtained on the Stanford-Binet for samples homogeneous in chronological age of which Jensen [illegible] aware, as it is mentioned in a later chapter, but its significance escapes him here. In the bivariate distribution of scores on the two forms of the 1937 test for groups homogeneous in chronological age the arrays are not homoscedastic, i.e., the scores of low-scoring examinees are closer together than those of high scorers. Jensen suggests a criterion for an interval scale that would reject this scale of the normally distributed Stanford-Binet. A transformation that would equalize the variances of the arrays would skew the distribution of total scores. It would have been better if this chapter had been omit-

ted. The distribution of scores has nothing to do with the definition of test bias used by the author.

The second error is indicated by the title of Chapter 6. It asks whether IQ tests really measure intelligence. For one thing this is all too close to the folk concept of real intelligence. A more important question is whether IQ tests measure anything useful and important, but this is answered quite satisfactorily in Chapter Eight. The specific problem the author is concerned with is the heterogeneous nature of the items in standard intelligence tests. He discusses this problem in terms of correlations and factors and gives his readers an introduction to these methods. Uninformed critics have frequently attacked intelligence tests on grounds of supposedly meaningless heterogeneity of content--a typical term used is hodge-podge--and have quoted authors such as Thurstone and Guilford, group factor theorists, in support of their criticisms. That very heterogeneous types of items are positively correlated with each other is important. That scores on a wide variety of cogritive tests are positively intercorrelated is also important. These positive correlations not only allow the use of heterogeneous items but require that use if the test is truly to measure the construct of general intelligence. If factor analysis is to add anything conceptually to the basic phenomena of

positive correlations among cognitive measures, the theorist must look beyond either the first principal component or the first principal factor, the definitions of the general factor used by the author.

Chapters 9, 10 and 11 contain the core of the methods and data concerning test bias as defined in the earlier chapter. If that definition is not accepted, there are neither methodologies nor data that can be brought to bear on the question. The first chapter in this group presents the methodologies for the study of bias based on the psychometric properties of the test studied in isolation and in relationship to external criteria. The second chapter presents the empirical data concerning correlations and regressions of total scores and criteria, and the third contains the empirical information about internal evidence of bias. These chapters will be discussed more fully later.

In Chapter 12 Jensen discusses the research relating the conditions of test administration to test bias. Although there have been many claims that blacks are handicapped by a wide variety of conditions of test administration, these amount to relatively little either singly or in toto. In an important sense this discussion also represents an anticlimax. Once it is established that test scores tend to overpredict black performance on socially important crite-

ria, which I shall discuss at greater length in the next section, it becomes obvious that test administration factors cannot amount to very much. Whatever the causes of the mean black deficit on mental tests may be, these causes also affect criterion performances.

Chapter 13 on sex bias is excellent. He discusses thoroughly the attempts to minimize sex differences in total scores on intelligence tests and the logical rationale for these efforts. Sex differences are ubiquitous given large enough samples. Differences occur not only on many item types but on individual items within item types, but such differences occur in both directions with approximately equal frequency. There is, of course, no defined population of items or item types from which samples could be drawn in order to obtain a more precise answer statistically. One wishes that he had confronted head on the common complaint that the same procedures should have been followed for blacks. There is a very simple answer to this complaint. One cannot find items having the properties of intelligence test items on which blacks equal or exceed the performance of whites.

Chapter 14 discusses comprehensively so-called culture-reduced techniques. This compulsive reviewer would have preferred the term "alternative" to "culture-reduced," but the choice

of term does not reduce the impact of the findings. An alternative that would reduce the difference between the means of blacks and whites to zero and would still perform as an intelligence test simply does not exist. In justification of my compulsiveness I doubt that tests such as those by Raven and Cattell are culture-reduced. The same is true of Piagetian tasks. These are all good measures of the general factor in intelligence, but they should be described as verbal-reduced. To the extent that tests are actually culture-reduced, such as measures of complex reaction time to culturally neutral stimuli, they are not very good measures of intelligence. Laboratory tasks can and do tell us a good deal theoretically about the nature of intelligence, but there is no replacement for standard intelligence tests in sight from this source.

The final chapter on the uses and abuses of tests is objective and dispassionate. If anything, the author can be charged with leaning over backwards with respect to his recommendations concerning proper uses of intelligence tests. Jensen applauds mainstreaming, but I am skeptical. Mainstreaming is not the only alternative to rigid tracking. I do not believe that the average teacher in an average classroom can meet the needs of children who represent the entire range of individual differences in readiness to do academic work at a given grade level.

He also states that there is very little reason to use intelligence tests in the early grades. Again I am skeptical. There is no reason to use intelligence tests at the time of school entrance merely to measure intelligence, but the test does assess the cognitive skills that have been acquired in the family and neighborhood. Learning at any level must start from a child's present skill and information if optimum results are to be achieved. An intelligence test can contribute to the design of an effective learning process.

THE PROBLEM OF TEST BIAS

Item bias occurs when different information is furnished by an item for two groups of examinees and many persons assume that such items are very common. A common criticism is that a black child could not possibly be expected to know the answer to a particular item or items. Jensen describes a number of different methods for answering the question of item bias and presents data obtained from the use of all of them. Some of these methods are a good deal more powerful than others, but even the least powerful is able to answer the common criticism with ease. There are very few items that stand out from others in furnishing different information about racial groups. If an item is difficult for a black child, the chances are very high that it will be difficult for a white child

as well.

It is necessary, as the author points out frequently throughout the entire book as well as in these chapters, to distinguish between statistical significance and social significance of a difference. It is reasonable to expect that there are no zero differences in nature. There will always be some small effect and, if samples are large enough, a statistically significant difference can be found. It is essential in all psychological research to define the limits of trivial effects. Studies of item bias are not an exception.

The author discusses item bias techniques in more or less the order in which they have appeared in the literature. There are correlations between distributions of item difficulties and analyses of variance to investigate the race by items interaction. Lack of item bias is shown by correlations between item difficulties for two groups approaching unity and by absence of interaction in analyses of variance. Both methods were later improved by transforming the scale of item difficulty proportions into a scale independent of the floor and ceiling effects of the raw proportions that attenuate the correlations and produce spurious interactions. In these investigations the evidence for item bias is small. Jensen then proceeds to show that most of the bias disappears when white sam-

ples are equated with black samples on the basis of total score on the test. This is readily done without regression effects by comparing younger whites with older blacks. Now if the same younger white samples are compared with older whites of the same age level as the blacks, evidence for an appreciable amount of bias returns, i.e., a difference in mean mental age produces evidence for some degree of item bias in all white population. There is, therefore, reason to believe overall ability level rather than cultural bias is responsible for what little bias is revealed by the first analyses.

As a matter of fact the small size of the race by items interaction is more than a bit surprising. It suggests that methods described up to this point are not very powerful although they are sufficient to reject the arm-chair arguments. If items differ in their validity as measures of the general factor in intelligence, and if blacks differ from whites on the general factor, there must be interactions between items and race even when mean mental ages are equal. Blacks and whites differ less on repeating digits than on other items on the Stanford-Binet. Such items when analyzed with verbal comprehension items produce interactions. This does not mean that all items on the test other than repeating digits should be discarded. The latter are among the least valid items. Some modest

amount of interaction is required for a standard intelligence test composed of heterogeneous items.

The distinction between statistical and social significance also leads to a conclusion that one cannot discard all items that show some evidence of bias. With respect to sex differences that seem to go in one direction about as frequently as they go in the other this means that discriminatory items should be balanced for the two directions. This balancing criterion might be applied, however, only to items within the limits of one's specification of trivial.

The balancing tactic cannot be applied directly when two groups differ in total score on the test. As described earlier it is possible to select a younger white group whose mean on total score is at the level of the black mean. Now one can apply the balancing criterion. For adults a white group could be matched with a black group on estimated true scores, and the number of items with a small amount of bias could be balanced. A colleague (Birnbaum, 1980) has suggested that evidence for bias be sought after the general factor has been held constant in the group differences on items. In any event test constructers cannot be purists and discard all items showing evidence of bias.

Jensen also describes the use of models de-

riving from latent trait theory as a methodology for testing for item bias. Lord (1980) has an all too brief chapter concerning this application as well. The approach would seem to be the most powerful method for testing for item bias, but there is one possibly important drawback. These models were developed on the assumption that the test is unidimensional but the best intelligence tests are quite heterogeneous with different clusters of items measuring several different factors. When the general factor is held constant, these group factors tend to be quite small, and the test can be considered unidimensional with respect to the major dimension. This is not the same, however, as the unidimensional test the theorists have in mind. It is just not known at the present time how far the assumption of unidimensionality can be relaxed in order to place examinees along the latent trait of general intelligence. It also goes without saying that racial comparisons involving these models on any tests have to be interpreted with respect to the social significance of the differences. With the sample sizes available to organizations like the Educational Testing Service it would be possible to show bias in some degree on most test items. As with sex differences one might impose a criterion of the size necessary to be nontrivial, discard the items that exceed this criterion, and balance the remaining items for the groups being compared.

The author reports other investigations directed at internal bias but none is as convincing to psychologically unsophisticated readers as those just described. Similar degrees of internal consistency, similar reliabilities, and other similar functional properties of tests for different groups all indicate little bias. If a test has highly similar psychometric properties other than a difference in means, it is not biased. One can conclude with considerable confidence that the evidence of test bias with respect to blacks is very meager indeed, whereas evidence for bias in tests for Hispanics occurs more frequently. A good bit of space is also given over to a discussion of the "Spearman hypothesis": that items and tests most highly loaded on the general factor will also discriminate most accurately between blacks and whites. The evidence in support of this hypothesis is not as strong as Jensen believes it to be--but more on this subject later.

Bias in predictive validation is also handled in a largely historical sequence. There was a rash of studies in which correlations between tests and criteria were compared with the null hypothesis of zero correlation in each group. This became known eventually as single group validation. Even if sample sizes were equal the independent significance tests would have been inadequate to reject the null hypothesis of a zero difference in the sample corre-

lations, but when the sample size for the minority group was substantially smaller than that for the majority group, an almost inevitable occurrence, one can only view this record with dismay.

There is no doubt that the comparison of regressions of criteria on predictor tests is required to evaluate bias in predictive validity, but psychologists were very slow in making use of these methods. Gulliksen and Wilks (1950) published the methods for making these comparisons about 20 years before they came into modestly general use. Shortly thereafter Humphreys (1952) recommended that the techniques be used in the study of test bias, and Gordon (1953) applied the techniques to the analysis of possible bias in Air Force Classification tests.

Gordon's data showed the same findings that have been found predominantly in recent years: namely, predictor tests most frequently over estimate black performance and underestimate white performance when a common regression equation is used for both groups but differences are not large. Inferences from test scores about socially important criterion performances are frequently in error, and the error is related to race, but the direction of the error is contrary to the expectations of the critics. One can also conclude from these data that the primary barrier for blacks in achieving proportional

representation in higher education, business, the professions, etc. is not the selection test but the criterion performance.

It is easy in reviewing this literature to lose the forest for the trees. In order to conclude that the null hypothesis of no difference in intercepts between regressions of two groups can be rejected one must have previously rejected the null hypotheses of no difference in error of estimate and no difference in slopes, in that order. A common cause of heterogeneity of errors of estimate is a floor on the test too high for the lower scoring group or a ceiling on the test too low for the higher scoring group. If one of the groups has been convinced by propaganda that the test is worthless or unfair or biased for them, these attitudes will produce a large error of estimate, low slope, and a large difference between group means but one can still plot the regression lines. None of these factors will produce a difference in performance on the criterion. Similarly a difference in slope does not require that the test be discarded or even that a single regression line will result in inferences contrary to the best interests of the minority group members. If the two regression lines cross, as indeed they must, three standard deviations below the majority group mean, there is no problem. As a matter of fact the crossing does not need to be at a point in the majority distribution that extreme. If the

crossing occurs well below the cut-off used for selection purposes, there is effectively no bias against the minority. It is also necessary in all of these comparisons to keep in mind the difference between statistical and social significance of differences. Given a large enough sample all hypotheses phrased as zero differences can be rejected.

There is one important point about regressions of criteria on predictor tests that Jensen misinterprets for his readers. Given classical scales of measurement it is true that the reliability of the criterion does not affect the slope of the regression line. However, the predicted criterion score is a very different score when the reliability of the criterion is increased. In the former condition the predicted score is one that on average contains a great deal of measurement error. In the latter condition the error component is much smaller. Furthermore, if the criterion measure is a test, or more generally a linear composite of several components, the slope of the regression line will also change with changes in reliability. The difference between the groups on the criterion, if the criterion is a test, will also increase as reliability increases.

The discussion of the reliability and validity of the predictor tests is also complex and is discussed in a completely adequate fash-

ion. As the reliability of the predictor increases, it is entirely possible that the inferences drawn from the test might change from overprediction of minority performance on the criterion, to accurate prediction, and finally to underprediction. In similar fashion the composite of two unbiased tests each measuring different components of criterion variance may prove to be biased. Since perfect lack of bias is impossible for two groups exhibiting a mean difference, it is essential that common sense be applied to the discussion of test bias. It is always possible to use separate regression equations, but the evidence to date is that an expectation of equal criterion performance of two groups would make the use of separate equations a more serious barrier to the selection of blacks than the use of a single equation.

Jensen also discusses in these chapters differences in definitions of culture bias, cultural loading, and fairness of test use. All tests are culturally loaded. As I have indicated earlier, culture reduced for some tests may assume too much. All socially relevant criteria are culturally loaded. If blacks, for example, are to achieve proportional representation in higher education, business, and the professions, they must be able to perform in the culture. After a discussion of test fairness that includes the attempts to define the term psychometrically, Jensen opts for the Peterson

and Novick (1976) position. Fairness is a matter of values. There is no single acceptable psychometric definition. There may certainly be a social concern that overrides the evidence of a predictor test in accepting larger numbers of blacks in higher education, etc., but this social concern should be valued in its own right and not used to attack the tests. Test bias can and should be studied and evaluated in a technical fashion.

GENERAL ISSUES

My first and primary issue is that Jensen has a predisposition to propose and stress genetic explanations to the exclusion of others. This predisposition, furthermore, detracts from the impact of his discussion of test bias. Although he announces in the Preface that he has excluded the genetic hypothesis for the difference in intelligence test score means for blacks and whites from his discussion of bias, it creeps in by inference if not by direct statement. With some reason Bronfenbrenner (1980) has called this unfair. For example, he refers several times to the heritability of intelligence within groups and typically uses a figure for h^2 of .70. He also tells the reader that this is quite a firm figure and that it is certainly no less than .50. Jensen's peers are somewhat more cautious (Loehlin, Lindzey, and Spuhler, 1975). Although I am not a peer of

Jensen in my knowledge of genetics, I am able to distinguish between well-controlled and poorly-controlled evidence. Family relationship correlations in homosapiens for finger print characteristics are satisfactory as evidence. Family relationship correlations for intelligence in which the genetic and environmental substrates for the development of intelligence are completely confounded are not satisfactory evidence. Given a large enough sample of children adopted at random into a wide enough range of adoptive families, correlations involving good measures of intelligence between both biological parents and both adoptive parents with the children would represent dependable data. A large enough sample of identical twins also adopted at random into a wide enough range of adoptive families would also be acceptable. I know of no data meeting these specifications.

There are other cases in point. I have earlier discussed the attempt to use the approximately normal distribution of scores on some intelligence tests as evidence in support of the polygenic hypothesis for intelligence. In his discussion of the relationship of intelligence to various measures of physique the only hypothesis considered is the genetic one. A seemingly obvious alternative is an environment factor or complex of factors of nutrition. For the explanation of the small correlation in adults of homogeneous racial stock of about .20

to .25 for intelligence and height he suggests a purely hypothetical explanation based on mutual assortative mating for height and intelligence to produce a between-families correlation. It is known that there is no genetic basis for a within-families correlation. The assortative mating explanation requires rather substantial correlations between husband's height and wife's intelligence, and the reverse, in addition to the known correlations of .50 for intelligence and .30 for height. I happen to be engaged at the moment in analyzing data concerning height and physique in which the hypothetical genetic explanation can explain only a small part of the relationships.

Yet another example of the predisposition to genetic explanations of various phenomena is his model for the observed degree of stability, or instability, of intelligence test scores over the years from preschool to adulthood. His model assumes a factor, genetic potential, that increases from one age level to another but on which individuals do not change in rank order. The instability is produced by random environmental factors. He states that this is the simplest model that will explain the stability data and at the same time explain the increasing parent-child correlation with increasing age of the child. I doubt that he has actually tried to quantify this model and fit it to the empirical observations.

The contribution to variance of individuals of the genetic substrate for intelligence is a problem orthogonal to the problem of bias in mental testing. It is not the only explanation for the data that relate the phenotypic measure of intelligence with other phenotypic measures, including other psychological tests, measures of physique, and measures of socially important criterion performances.

The second issue is his account of the various tests of the so-called Spearman hypothesis and the interpretation of the findings. The hypothesis states that the types of items measuring "g" most validly are also the items that discriminate between blacks and whites most sharply. Is a statistically significant correlation, whatever that may mean in these circumstances, between the two distributions of correlations sufficient to discard the null hypothesis? There is no population of test items or item types from which random samples can be drawn. Neither are the units of sampling independent of each other. The investigator is forced to rely on a descriptive statistic of the joint distributions. Strong confirmation would be found if the data points formed something very close to a straight line and if each test variable entering the respective correlations were equally reliable. From the size of the correlations, frequently based on tests or items having very different reliabilities, con-

firmation of the hypothesis is weak at best. I have stressed the need for equal reliabilities because differential amounts of measurement error spuriously increase the size of the correlations reported.

Humphreys, Fleishman, and Lin (1977) found that when whites were dichotomized into low and high SES groups on the distribution of individual SES scores, and means were obtained on Project Talent tests with sex, section of the country, and age taken into account, the profiles of the two groups became almost mirror images of each other. Eight comparisons of groups differing in sex, area of the country, grade in high school, and socio economic status had correlations between profiles ranging from -.87 to -.93. Unfortunately, there was some contribution to the mirror image relationship by differential reliability of the measures in the profile. When the same comparison is made between black and low SES profiles, in spite of differential reliability, the two profiles are almost independent of each other. There may be an overall difference in "g" between blacks and low SES whites, since the across-the-board means do differ, but there are other causes of the size of the differences on individual tests and thus of the across-the-board difference that are independent of race. It is quite certain that the Spearman hypothesis is given more support in white social class comparisons than in

race comparisons in a very heterogeneous selection of tests. It is also probable that I have devoted too much space to this matter since it is really quite independent of the problem of test bias. A test could be badly biased even if a large portion of its variance were determined genetically, and a test could be unbiased even if its variance were almost entirely environmental in origin.

The discussion of test theory is a third area in which I take issue with the author--and with a good many test theorists as well. Jensen takes the position that it is necessary to compromise among the features of a psychometrically perfect test to construct an intelligence test. He and others are wedded to the concept of a factor-pure test or a unidimensional highly homogeneous test. Jensen even defines reliability in terms of homogeneity or internal consistency. The description of a factor pure test is phrased in statistical terms with an unspecified assumption that statistical homogeneity is the same as psychological homogeneity.

It is possible to question these underpinnings of present day test theory (Humphreys, 1970). For example, Guilford would call a test of any of his 120 basic abilities a factor-pure test. These abilities are defined, however, as the Cartesian products of three facets containing 6, 5, and 4 elements, respectively. Each

factor-pure test is inextricably complex psychologically since each measures simultaneously one element each of content, operation, and product. Note that it is impossible to construct a measure of any element without contamination of elements on other facets, but it is possible to vary these other elements so widely that a relatively pure test of general intelligence can be constructed with the confounding variance reduced to tiny amounts by making the test as heterogeneous as possible within the limits defined by the attribute one wishes to measure.

From this point of view one makes no compromises in constructing a valid test of general intelligence. It is also necessary from this point of view to distinguish sharply between reliability and homogeneity. The reliability of the total score on a psychological test is hopelessly confounded with the homogeneity of the items and if the test constructor is not careful validity can be sacrificed on the altar of homogeneity. Reliability is best defined by the correlation between test and retest. If retest is barred by the memory factor, a parallel form is acceptable. This should still be the definition of reliability even in those few cases psychologically in which neither test-retest nor parallel form is readily apparent, e.g., the number of four-letter words starting with S that an examinee can utter in a fixed time period.

Jensen quotes early in his development of general factor theory the very important principle of Spearman: namely, the indifference of the indicator. There are many sources of items. Good items are independent of the content, operation, or product. Items that measure eduction of relations and correlates are good measures of intelligence, but so are many, many other items. It is possible to develop an intelligence test entirely from information items as long as the kinds of information called for are sufficiently heterogeneous. It also follows that Ravens' matrices are not pure measures of general intelligence. These items carry their own excess baggage of unwanted contributions to variance. It may be possible to develop a good test of intelligence with fewer items if the test constructor concentrates on measuring education of relations and correlates rather than on information items, but the former items would have to include more than figural content and involve many operations.

A fourth area of general concern is statistical. This manuscript should have been read very carefully by a competent statistician before publication. I shall not attempt to list all of the errors I found. For the most part, as stated earlier, they do not detract from the aim of the book. Reviewers who concentrate on these may well influence readers to underestimate drastically the importance of chapters 9,

10, and 11, but some errors must be discussed.

On two occasions Jensen misinterprets partial correlations. The first of these occurs in an early chapter in which he analyzes the variance of scores on intelligence tests into components based on race, socio-economic status, between families within race and SES groups, and within families within race and SES groups. He computes partial correlations between race and intelligence with SES held constant and between SES and intelligence with race held constant, squares these, adds them, and divides the residual variance into the between and within families components. This error resulted in an underestimate of the variance contributed by race and SES of about 8%. He should, of course, have squared the multiple correlation of intelligence with race and SES. The remainder is shared by between, within, and error sources. The squared multiple correlation can in turn be broken down into two squared part correlations for race and SES, respectively, plus their joint contribution.

The presentations of analyses of variance are defective. In a design involving between and within subject variation it is misleading to report contributions to total variance. The variance associated with within subjects variables is constrained by the dependencies among the levels of treatment. It is better to sepa-

rate the two different categories of variance in one's interpretations. More importantly eta squared is a squared correlation between two variables. The eta squared for race and intelligence requires use of the between subjects contributions to variance only. If there are additional between subjects variables, a correlation that excludes all independent contributions to variance except the one under consideration and the contributor of error is actually a partial correlation. For the within subjects set of variables and interactions each eta squared requires an appropriate error term if it is to be interpreted accurately. Finally, it is clear that an error term cannot itself be described by an eta squared.

One last methodological issue requires comment. Jensen recognizes the difference between the first principal component and the first principal factor, but appears to use either without discrimination as his operational definition of the general factor. The former is an especially unfortunate choice when one or more columns of correlations are all small. However, neither is adequate since both convert small or even zero correlations into negative first factor residuals. It is logically inacceptable that a general factor can be defined by zero correlations or that partialing out the general factor from a correlation between two measures should result in a negative residual.

Even though it may frequently be difficult computationally to obtain a more defensible general factor, the author should recognize that possibility.

CONCLUSION

As I look back on this review it appears more critical than it should be. The book was a very difficult undertaking, and with certain exceptions that bear relatively little on the principal thrust of the discussion it was well done. It is important to state that the data presented are generally sound, and those that bear most directly on test bias cannot be dismissed. The problem of ethnic differences in phenotypic intelligence is also a pressing one in our society. This is especially true for the difference between blacks and whites. The emotions with which it is charged make almost impossible rational consideration of the issues, yet those issues must be faced in all of their ramifications. To take a dramatic example, an important problem underlying the discussion of the all-volunteer army is the low level of scores on qualifying and classification tests, and the borderline performance in military service, of a large proportion of enlistees. Given the increasing proportion of blacks in military service, the problem is largely though not entirely a black problem. Secretary of the Army

Alexander is so identified with the problems of blacks that his only recourse is to make irrational emotional criticisms of the tests used in military selection (Holden, 1980) that insult his intelligence as well as ours. He is so caught up emotionally in this issue that he is jeopardizing to some degree our national security.

The implications of some little noticed conclusions of Jencks et al (1972), also quoted by Jensen, are of considerable importance in this regard. A system in which standardized tests and academic grades were the sole determiners of educational and occupational advantage would substantially benefit working class white students relative to their middle class counterparts. The present advantage of the latter would be reduced by at least a third. This system would, however, be far less satisfactory for blacks than for poor whites. I quote the following (p. 195): "For blacks, the ideal system is one which discounts test scores and emphasizes aspirations. Failing this, a system that emphasizes credentials is better for blacks than one that emphasizes test scores. The right policy thus depends on whose opportunities you want to equalize." This conflict fuels a good deal of the criticism of tests of mental ability. For a different subset of our population it fuels antagonism toward strong affirmative action policies based on rigid goals or quotas.

In my judgment the available data on test bias cannot be rejected as unimportant or irrelevant. The deficits in academic skills of black youth at the time of leaving the public schools effectively bar large numbers from higher education and occupations of the highest prestige in our society. The effect of these deficits at the lower end of the hierarchy of occupational prestige or in military service should also not be underestimated. Consider the following list of academic skills and knowledge: aural and visual comprehension of standard English, writing of standard English, arithmetic, introductory mathematics, information about science, social studies, and industrial and domestic arts. Now consider, given reliable measures of each of these, that the following approximate percentages of black youth at 18 fall at and below the grade level indicated: 50% at grade 9, 15-20% at grade 6, 2-3% at grade 3. Even a relatively low level occupational skill such as typing speed and accuracy depends on the typist being able to type phrases and clauses as a unit. This cannot be done if the typist cannot spell, capitalize, and punctuate and does not understand word order in sentences. I have yet to see good evidence that much compensation for academic deficiencies occurs after age 18. Black students admitted to college with borderline academic credentials also enter graduate and professional schools with borderline academic credentials. They later fail objective

professional licensing examinations in disproportionate numbers.

The above does not mean that compensation is impossible. Neither does it mean that the academic deficiencies are caused by low genetic capacity measured by intelligence tests. Intelligence is a phenotypic trait, and a composite formed from tests measuring each of the academic skills and types of knowledge listed above is also a very good measure of that same trait. The rather different content of intelligence and broad-band academic achievement tests is immaterial when young people have been exposed to the public school curriculum for 12 years.

The black deficit in intelligence has important personal consequences for black children and important social consequences for all of us. Abolishing the tests will do nothing to abolish the deficit. Obtaining credentials in greater numbers, such as high school diplomas and college degrees, will not abolish the deficit unless these credentials are associated with increased learning. We may not know enough presently to design effective programs to promote intellectual gains, but we shall not acquire the necessary knowledge by refusing to recognize the problem. We shall also not recognize the problem if the cry of "racism" is raised every time an attempt is made to analyze and describe it.

References

Bronfenbrenner, U. Intelligence in black and white. Washington Post, Sunday, April 20, 1980.

Gordon, M. A study in the applicability of the same minimum qualifying scores for technical schools to white males, WAF, and Negro males. Technical Report 53-34, Lackland AFB, Texas, Personnel Research Laboratory, 1953.

Grady, J. PASE vs. Hannon. Federal District Court, Chicago, 1980.

Gulliksen, H., and Wilks, S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Holden, C. Doubts mounting about all-volunteer force. Science, 1980, 209, 1096.

Humphreys, L. Individual differences. Annual Review of Psychology, 1952, 3, 131-150.

Humphreys, L. A skeptical look at the factor pure test. In Current Problems and Techniques in Multivariate Psychology. C. E. Lunneborg (ed.) The University of Washington, Seattle, 1970, 23-32.

Humphreys, L., Fleishman, A., and Lin, P. Causes of racial and socio-economic differences in cognitive tests. Journal of Research in Personality, 1977, 11, 191-208.

Jencks, C. Inequality: a reassessment of the effect of family and schooling in America. Basic Books, New York, 1972, 399 pp.

Loehlin, J., Lindzey, G., and Spuhler, J. Race differences in intelligence. Freeman and Co., San Francisco, 1975, 380 pp.

Lord, F. Applications of item response theory to practical testing problems. Erlbaum, New York, 1980, 274 pp.

Peterson, N. and Novick, M. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Note

Birnbaum, M. Personal communication. 1980.